17.02.2023, 15:58 HW2 Парсинг

## Урок 2. Парсинг данных. HTML, DOM, XPath

Написать приложение и функцию, которые собирают основные новости с сайта на выбор dzen.ru, lenta.ru, mail.ru . Для парсинга использовать XPath

Структура данных должна содержать:

- название источника
- наименование новости
- ссылку на новость
- дата публикации

```
In [1]:
         from pprint import pprint
         from lxml import html
In [2]:
         import requests
In [3]:
In [12]:
         url = 'https://news.mail.ru/'
         headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/53
In [13]:
         response = requests.get(url, headers=headers)
         dom = html.fromstring(response.text)
In [32]:
         news_list = []
         news_elements = dom.xpath("//div[@class='newsitem newsitem_height_fixed js-ago-wrap
         news_elements
         [<Element div at 0x24eeb8ba900>,
Out[32]:
          <Element div at 0x24eeb8bae50>,
          <Element div at 0x24eecdcac70>,
          <Element div at 0x24eecdcae00>]
In [34]: for element in news_elements:
             herf = element.xpath(".//span[@class='cell']/a/@href")
             title = element.xpath(".//span[@class='cell']/a/span/text()")
             source = element.xpath(".//div[@class='newsitem__params']/span[@class='newsiter
             time_pub = element.xpath(".//div[@class='newsitem__params']/span[@class='newsit
             print(f'Новость - {title}\nИсточник - {source}\nСсылка - {herf}\nВремя публика
```

17.02.2023, 15:58 HW2 Парсинг

```
Новость - ['В КНДР пообещали беспрецедентный ответ на учения США и Южной Кореи']
Источник - ['Коммерсантъ']
Ссылка - ['https://news.mail.ru/politics/55098117/']
Время публикации - ['2023-02-17Т09:57:55+03:00']
Новость - ['ЦБ РФ с 1 апреля запустит пилотный проект по операциям с цифровым рубл
ем с 13 банками']
Источник - ['ТАСС']
Ссылка - ['https://news.mail.ru/economics/55100425/']
Время публикации - ['2023-02-17Т11:50:49+03:00']
Новость - ['Исследование: что происходило с температурой Земли до глобального поте
пления']
Источник - ['Новости Mail.ru']
Ссылка - ['https://pogoda.mail.ru/news/55101108/']
Время публикации - ['2023-02-17T12:19:43+03:00']
Новость - ['В России впервые с 13 октября 2022 года выявили более 15 тыс. заболевш
их ковидом за сутки']
Источник - ['ТАСС']
Ссылка - ['https://news.mail.ru/incident/55099929/']
Время публикации - ['2023-02-17T12:37:36+03:00']
```

In [ ]: